The Journal of Experimental Education

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: https://www.tandfonline.com/loi/vjxe20

Latent Class Evaluation in Educational Trials: What Percentage of Children Benefits from an Intervention?

Germaine Uwimpuhwe, Akansha Singh, Steve Higgins, Mickael Coux, ZhiMin Xiao, Ziv Shkedy & Adetayo Kasim

To cite this article: Germaine Uwimpuhwe, Akansha Singh, Steve Higgins, Mickael Coux, ZhiMin Xiao, Ziv Shkedy & Adetayo Kasim (2020): Latent Class Evaluation in Educational Trials: What Percentage of Children Benefits from an Intervention?, The Journal of Experimental Education, DOI: 10.1080/00220973.2020.1767021

To link to this article: https://doi.org/10.1080/00220973.2020.1767021

THE JOURNAL OF EXPERIMENTAL EDUCATION
https://doi.org/10.1080/00220973.2020.1767021

Routledge
Taylor & Francis Group

MEASUREMENT, STATISTICS, AND RESEARCH DESIGN

Latent Class Evaluation in Educational Trials: What Percentage of Children Benefits from an Intervention?

Germaine Uwimpuhwe[a], Akansha Singh[a], Steve Higgins[a], Mickael Coux[a], ZhiMin Xiao[b], Ziv Shkedy[c], and Adetayo Kasim[a]

[a]Durham University, UK; [b]University of Exeter, UK; [c]University of Hasselt, Belgium

ABSTRACT

Educational stakeholders are keen to know the magnitude and importance of different interventions. However, the way evidence is communicated to support understanding of the effectiveness of an intervention is controversial. Typically studies in education have used the standardised mean difference as a measure of the impact of interventions. This measure, commonly known as the effect size, is problematic, in terms of how it is interpreted and understood. In this study, we propose a "gain index" as an alternative metric for quantifying and communicating the effectiveness of an intervention. This is estimated as the difference in the percentage of children who make positive gains between the intervention and control groups. Analysis of four randomized controlled trials in education supports the expectation that most children make progress due to normal school activities, which is independent of the intervention. This study elaborates a method to illustrate how trials with a positive gain index and with a higher percentage of pupils with positive gain in the intervention group can be used to communicate which trials are effective in improving educational outcomes.

KEYWORDS

Gain index; effect size; mixture model; bayesian approach

Introduction

RANDOMIZED CONTROLLED TRIALS (RCTs) are the most powerful tool used in the hierarchy of evidence to establish a cause-effect relationship between an intervention and an outcome. The method provides a robust technique to measure the effectiveness of an intervention (Hutchison & Styles, 2010). In educational trials, pupils are randomly assigned either to receive an intervention or to continue with the usual schooling program (Hariton & Locascio, 2018; Kendall, 2003). Each school or pupil has the same chance of being allocated into one of the two groups (in a two parallel arms trial with equal allocation ratio and depending on whether it is an individual or cluster-randomized trial). Randomization, if done properly, can help to eliminate selection bias (Torgerson & Torgerson, 2013) and provide a more accurate estimate of the effect of an intervention. In this paper we explore some of the challenges associated with using effect sizes as a measure of impact, we then summarize some of the alternative ways that have been suggested to communicate this effect, before describing a method which we believe addresses the major challenges and avoids many of the issues associated with alternative proposals.

CONTACT Adetayo Kasim a.s.kasim@durham.ac.uk Department of Anthropology, Durham University, Durham, UK.

Supplemental data for this article is available online at https://doi.org/10.1080/00220973.2020.1767021.

Challenges associated with the interpretation of effect sizes

The most common measure of impact used in education trials is a standardized mean difference and its associated confidence intervals, known as the effect size. It is a ratio measure, the difference in means between two groups, divided by the pooled standard deviation of these groups. Whilst there are different methods of calculation, it is usually reported as Cohen's d or Hedges' g (see Bakker et al., 2019 for more details about these calculations and alternatives). Hedges' g effect size is more suitable for a small number of participants than Cohen's d, but both metrics are equivalent for 30 or more number of participants. Although the impact of a particular intervention could be based on an unstandardized mean difference, this can only be used to compare effects across studies with the same outcome represented on the same scale (Vacha-Haase & Thompson, 2004). However, measurement outcomes are often scaled differently, so comparing unstandardized mean differences across studies is not meaningful or practical (Durlak, 2009). The main aim of standardization is therefore to equate effects measured on different scales to provide a comparable metric across similar measures. A highly desirable property of an effect size based on a standardized mean difference is the stability of estimates between different versions of the same instrument, individual scores, and different study designs (Baguley, 2009).

Teachers and other educational stakeholders who do not usually have a statistical background do not easily understand the metric of an effect size from a standardized mean difference. Ratio relationships and measures are particularly problematic (Izsák & Jacobson, 2017). As a result, Cohen (1969) proposed a proportional or percentile indices to aid in the interpretation of effect sizes based on the overlapping proportions of two normal distributions for two known groups by identifying these simply as small, medium or large (see also Rosenthal & Rubin, 1979, 1982). Cohen's categorization (Cohen, 1977, 1988) is commonly cited, especially in social and behavioral research. However, assuming universal thresholds for effect sizes poses its own challenges because the categories proposed by Cohen are arbitrary. Cohen himself suggested that his categorization should not be generalized for all effect sizes in different areas. Glass et al. (1981) argued that even a very small effect size of 0.10 can be considered important especially if its application is inexpensive or easy to achieve. More importantly, McCartney and Rosenthal (2000) also observed that Cohen's interpretation cannot be used to assess interventions in education where it is often difficult to change the trajectory of students' learning. A small effect may be particularly valuable. Cohen's categorization has been consistently criticized by many other researchers (Ferguson, 2009; Hedges & Hedberg, 2007; Hill et al., 2008; Lipsey et al., 2012). Recently, Funder and Ozer (2019) revised Cohen's thresholds using correlation as a measure of effect size. They defined an effect size of 0.05, 0.10, 0.20, 0.30, and 0.40 as very small, small, medium, large, and very large effect respectively. However, any generalization of thresholds is not a sound approach in education because of a number of factors which influence the measure, such as the type of intervention and the age of the participants (Hill et al., 2008). Vocabulary interventions, for example, tend to have larger effect sizes as the outcomes are relatively simple, reading comprehension estimates tend to be much lower, as this capability involves the co-ordination of a number of aspects of reading such decoding, vocabulary, grammar and syntax (Higgins, 2018).

There have been a number of other proposed alternative interpretations and representations of effect size. Rosenthal and Rubin (1982) suggested the binomial effect size display (BESD). This was designed to indicate the practical importance of any particular effect size estimate. It shows the difference between two proportions such as the difference between the success rate of a new intervention and the success rate of the standard intervention, based on assumptions about the underlying distributions. McGraw and Wong (1992) suggested a "Common Language Effect Size (CLES)" statistic, which is the probability that a score sampled at random from one distribution will be greater than a score sampled from another distribution. Coe (2002) interpreted effect sizes as z-scores and proposed looking at the percentage of pupils in the control group who are below an average pupil in the intervention group. Lipsey et al. (2012), in their review of the

Table 1. Summary statistics of the four trials used in this paper.

	Comparison		Intervention		Overall	
	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N
Pretest						
Trial 1	16.57 (6.84)	142	16.97 (7.74)	149	16.78 (7.3)	291
Trial 2	85.42 (11.18)	143	85.26 (10.67)	159	85.33 (10.9)	302
Trial 3	22.87 (5.25)	89	22.90 (5.24)	93	22.88 (5.23)	182
Trial 4	24.35 (4.27)	192	24.69 (4.29)	199	24.52 (4.28)	391
Post-test						
Trial 1	18.56 (8.13)	142	21.5 (8.08)	149	20.06 (8.22)	291
Trial 2	84.15 (11.35)	143	87.61 (10.2)	159	85.97 (10.88)	302
Trial 3	82.76 (9.96)	89	83.99 (10.01)	93	83.39 (9.98)	182
Trial 4	88.85 (10.88)	192	88.77 (11.55)	199	88.81 (11.21)	391

interpretation of effect sizes, concluded that appropriate norms based on distributions of effect sizes for comparable outcome measures from comparable interventions targeted on comparable samples need to be established to decide the magnitude of effect size as small, medium or large in a particular context.

Why the gain index?

Educational stakeholders such as policymakers, parents, and teachers do not easily understand these existing metrics (Baird & Pane, 2019; Rosenthal & Rubin, 1982). Yet the wider community and policymakers seem to have less difficulty understanding percentages presented in a 2×2 table (Randolph & Edmondson, 2005). This kind of presentation of results is similar to the BESD, which can be computed relatively easily to show the effect (Rosenthal, 1990; Rosenthal & Rosnow, 1991). BESD, however, relies on the underlying effect size and assumes the overall improvement rate is 50%, which is highly unlikely in most education trials. It is also only a broad approximation with substantial error (Miller et al., 2011).

The Education Endowment Foundation, adopted the idea of months of progress (Higgins et al., 2015), which converts effect size to additional progress in months attributable to an intervention, based on the assumption that one standard deviation is equivalent to one year of progress on standardized tests (Glass et al., 1981). Hill et al. (2008) challenged this assumption and argued that effect sizes vary according to the ages of the children and the tests used (Hill et al., 2008). Months of progress can therefore lead to unreasonable conclusions and misinterpretation (Baird & Pane, 2019). All of these challenges illustrate the central tension in terms of communicating effects which can be summarized as the tradeoff between accuracy and accessibility (Higgins, 2018). The effect size (with its associated confidence intervals) is accurate, but not very accessible. Approaches which create arbitrary thresholds or make assumptions about the underlying distributions have been shown to be more easily understood but can be inaccurate or misleading.

In this paper, we propose a new measure to evaluate interventions in education trials that can be interpreted easily by policymakers and educational stakeholders. Instead of using the standardized mean difference as an effect size with its inherent problems outlined above, the effectiveness of an educational intervention can be evaluated and communicated by simply estimating more precisely the percentages of children who make improvement beyond the expected level (positive gain) in the intervention and comparison groups, without making the distributional assumptions adopted in earlier measures. In the real world, students are expected to improve over time based on normal school teaching and learning activities, even if they do not receive any particular 'intervention'. The difference in percentages of positive gain between intervention and comparison groups, which we termed as the "gain index" provides useful and simple information about the percentages of the participating children that are likely to have benefited from the

intervention. This simple metric based on expressing the effectiveness of an intervention as percentages can therefore be a useful and intuitive tool for communicating evidence in educational trials to parents, schools, and policymakers. More importantly, it can be presented in a similar way to BESD. However, the ease of interpretation comes at the cost of a more computationally intensive method.

Data and methods

We analyzed eighteen randomized controlled trials funded by the Education Endowment Foundation in England. Only four of them are discussed in detail in the results section. An overview of these four trials is provided below. The overall summary statistics of the four trials such as mean, standard deviation and sample size are presented in Table 1.

Graduate coaching programme efficacy trial (trial 1)

The aim of Perry Beeches Graduate Coaching Program was to improve the reading and writing skills of Year 7 pupils with low levels of attainment in four English secondary schools. In this project, 16 graduate 'coaches' provided academic support to pupils who had not reached level 4c in English at the end of Key Stage 2, which is the expected level for 11 year old at the end of primary education. Originally, it was intended to provide one to one support to pupils from graduate coaches. However, in practice, pupils received a range of targeted support that varied between schools and not all coaches were graduates. The program built on a successful pilot in Perry Beeches Academy in Birmingham, and the school coordinated the project across participating schools. The approach was based on a one to one coaching program used in the Match Charter School in Boston, USA. The impact of the program was assessed on the academic outcomes of 291 students from 4 schools who were offered support during the 2013–2014 school year. The project was selected by the EEF as one of 24 projects in a themed round on literacy catch-up at the primary-secondary school transition. Projects funded within this round aimed to identify effective ways to support pupils not achieving level 4 in English at the end of Key Stage 2 and at the end of primary school. This is the expected level that 11 year old pupils are expected to achieve. The primary outcome was pupils' reading and writing ability, specifically reading, spelling and grammar as assessed by the GL Assessment Progress in English (PiE) Test Short Form paper version. More information about this trial can be found in Lord et al. (2015).

Butterfly phonics (trial 2)

Butterfly Phonics aimed to improve the reading of struggling pupils through phonics instruction and a formal teaching style where pupils sit at desks in rows facing the teacher.

The pupils who did not reach level 4 in their Key Stage 2 national tests or their reading skills were at least a year behind their chronological age were eligible to participate in the trial. The lessons were taught for ten to twelve weeks, typically with two one-hour lessons each week in schools. Pupils were taught in small groups of typically six to eight by a trained Butterfly practitioner and assisted by a trained teaching assistant. The randomized controlled trial evaluation consisted of a treatment group, which received the intervention, and a control group who continued their schooling as usual. The intervention took place in school time for five out of the six participating schools and in a variety of lessons. No control activities were involved so the comparison was 'business as usual'. The unit of randomization was the individual pupil and the primary outcome was reading comprehension. More information about this trial can be found in Merrell and Kasim (2015).

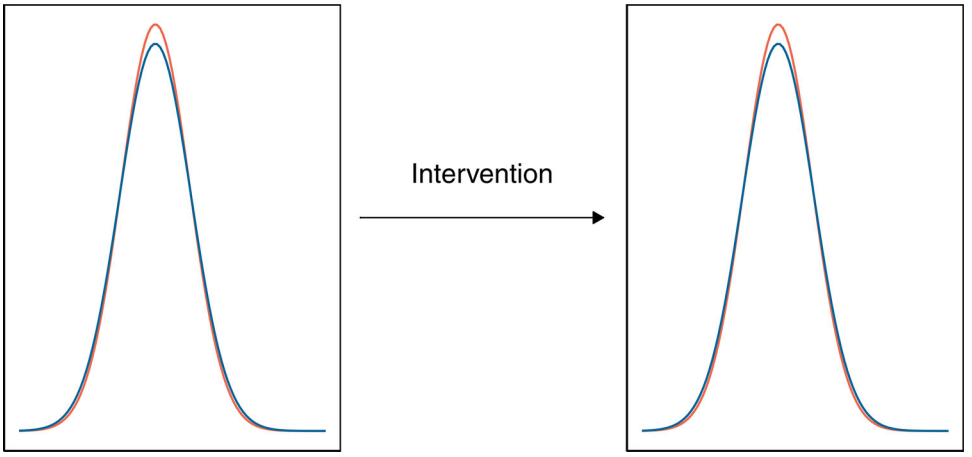

Figure 1. A simulated illustration of the expected distribution of gain scores between pre and post-intervention period when there is no intervention effect.

Summer active reading (trial 3)

The Summer Active Reading Program aimed to improve reading skills and particularly comprehension by raising pupil's engagement in, and enjoyment of, reading at the transition from primary school to secondary school. Booktrust, an independent charity that aims to change lives through engaging people with reading delivered this program. Four book packs were gifted to participating pupils, who were invited to attend two summer events led by Booktrust staff at their new secondary school. The first book pack was gifted toward the end of the child's final term at primary school, the second and third packs at two summer events and the final pack in the first term of secondary school. Volunteers, recruited by Booktrust, gifted the book packs and supported activities, including one to one reading, at the summer events. The trial examined the impact of the program on 205 pupils from 10 schools in the north of England who had been identified as unlikely to achieve Level 4a or above by the end of Key Stage 2 (the expected level of achievement for 11 year olds at the end of primary schooling). Pupils who were not likely to gain at least Level 2 (the expected level for 7 year olds) were not included in the trial. More information can be found in Maxwell, Connolly, Demack, O'Hare, Stevens, Clague, and Stiell (2014).

TextNow (trial 4)

The aim of TextNow Transition Program was to improve the reading comprehension skills of pupils at the transition from primary to secondary school. Unitas, a national charity that helps young people access, participate, and progress in mainstream education and training delivered the program. Participating students received 20-minute one to one sessions with a volunteer coach each weekday for five weeks at the end of primary school and for a further 10 weeks at the start of secondary school. Children were expected to read independently for a further 20 minutes per day and were rewarded for attendance with credits that could be used to buy books online. The impact of the program in this trial was assessed on 501 pupils in 96 schools across England who had been identified as unlikely to achieve Level 4a or above by the end of Key Stage 2, but have gained at least Level 2. More information can be found in Maxwell, Connolly, Demack, O'Hare, Stevens, and Clague (2014).

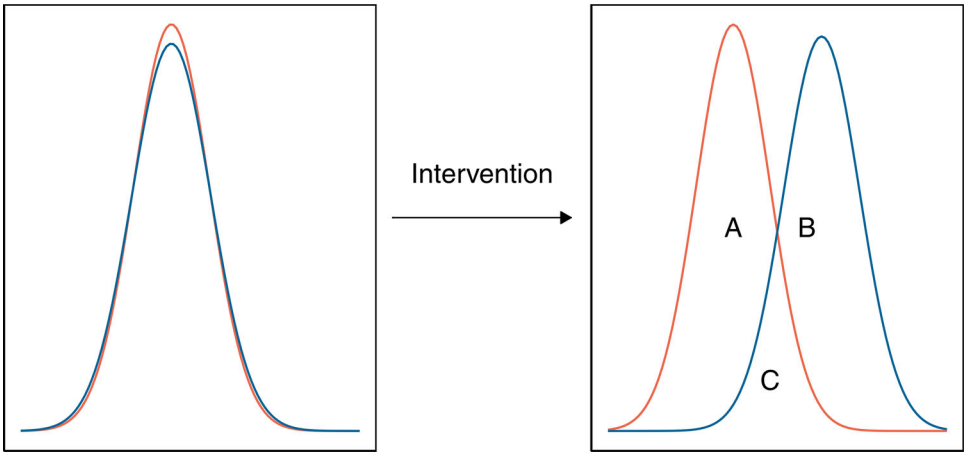

Figure 2. A simulated illustration of the expected distribution of children attainment between the pre and post-intervention period when there is a significant intervention effect.

Gain index

The educational outcomes for schoolchildren are expected to improve over time due to normal teaching and learning activities in school. This means that most students tend to make progress whether or not they participated in an education trial. However, it is also expected that children who participate in an educational intervention should make more progress than those in the comparison group if the educational intervention is effective. Figure 1 shows the hypothetical distribution of pretest and post-test scores for children randomized to hypothetical intervention and comparison groups when the education intervention has no beneficial effect. In this scenario, the distributions of pretest and post-test scores are the same during pre- and post-intervention period. However, the average of the post-test scores would usually be larger than the average of pretest scores due to progress resulting from teaching in school and other activities. Hypothetically, for an intervention with no beneficial effect, it can reasonably be expected that the percentage of children who make positive gains in the scenario as depicted in Figure 1 will be approximately the same in both the intervention and comparison groups.

Figure 2 presents a scenario where the educational intervention is assumed to be effective on average, and beneficial to the children who were randomized to the intervention group. The distribution of the test scores during the pre-intervention period is the same in both the intervention and the comparison groups. However, the distribution of the test score post-intervention in the intervention group is shifted further to the right than the distribution of the post-test score in the comparison groups. Usually, the impact of the intervention is calculated as the standardized mean difference between the two distributions or groups (the difference in the post-test scores, corrected for pretest and divided by the pooled standard deviation, or its equivalent from a regression analysis). But an interesting feature in Figure 2 is that a higher percentage of the children in the intervention group are more likely to have a higher post-test score than those in the comparison group. This means that the impact of intervention can be calculated as the percentage difference in the number of children that are more likely to have higher scores post-intervention than pre-intervention. This percentage difference is what we have called the "gain index".

Bayesian shared parameter mixture model

The concept of the "gain index" seems very straight forward, except it is important that the definition of children who are likely to have higher scores post-intervention than pre-intervention

needs to be clearly defined. This means that it is not sufficient to cluster children simply by positive or negative scores based on the difference between post-test and pretest scores. For example, a child who had +1 score is not necessarily different from another child with +1.5 score in this case. The difference between post-test and pretest scores depends on the uncertainty associated with each individual child's score.

To objectively define those who are more likely to have a higher score post-intervention than pre-intervention, we propose to assume that all children can be grouped based on an unobserved latent construct (Wedel, 2002). It consists of two groups of those that are likely to make positive gains (i.e. having higher scores post-intervention than pre-intervention) and those that are likely to make a loss (i.e. having lower scores post-intervention than pre-intervention: negative gains). Although the two groups are initially unknown, they can be determined based on the observed gain score data using a finite component mixture model. A similar approach has been applied in other areas of social science to explore unobserved groups. For instance in alcohol consumption among adolescents (Wang & Bodner, 2007), alcohol consumption and marijuana consumption (Hix-Small et al., 2004) and the effectiveness of learning during working memory for nursery and primary school children (Orylska et al., 2019). A mixture model is a flexible technique for identifying latent components by approximating the distribution of the observed data by a mixture of distributions (Geoffrey & Peel, 2000).

Let y_i represent the gain score for pupil i in a school. The gain score y_i is obtained as a residual from an ANCOVA model, where $i = 1, 2, ., N$ represents the number of pupils in the school. Assuming k subgroups of pupils in the school from the least performing group to the highest performing group, the finite component mixture model can be formulated as:

$$f(y_i) \sim \sum_{j=1}^{k} p_j f_j\left(y_i \mid \theta_j\right), \tag{1}$$

where $p_j \geq 0$ and $\sum_{j=1}^{k} p_j = 1$.

In the simplest situation, the distributions f_j's specified in Equation (1) are known and inference focuses either on the proportions p_j or on the allocations of the observation $y_1, ..., y_N$ to components f_j. In most cases, f_j's represent a parametric family with unknown parameter θ_j. When the focus is to determine which observation y_j belongs to which of the k components, as is the case here, the shared parameter mixture model can be formulated with latent variable representation. Furthermore, to account for the clustering of pupils nested within schools, we adopted the shared parameter mixture model proposed by Evans and Erlandson (2004) which can be formulated as:

$$f(y_{is}|b_s) = \sum_{j=1}^{k} z_{isj} f(\mu_j + b_s, \sigma_j^2). \tag{2}$$

where y_{is} is the score of pupil i from school s and $s = 1, 2, ..., S$. Here S is the total number of schools and $n_s = n_1, n_2, ..., n_S$, represents the number of pupils within schools. $z_{is} = 1$ if pupil i makes + ve gain and $z_{is} = 0$ if the pupil makes − ve gain, and $b_s \sim N(0, \sigma_b^2)$ is the random intercept for school s. Assuming a T-distribution and two components of positive and negative gains, the model defined in Equation (2) is re-formulated as:

$$f(y_{is}|b_s) = (1 - z_{is})\mathcal{T}(\mu_1 + b_s, \sigma^2, df) + z_{is}\mathcal{T}(\mu_2 + b_s, \sigma^2, df). \tag{3}$$

Note that $p_2 = \sum_{i=1}^{N} z_{is}/N$ is the true and unobserved probability that a child belongs to a + ve gain component (quality subgroup) in the population, $p_1 = 1 - p_2$ is the true and unobserved probability that a child belong to a − ve gain component in the population. Note that the shared parameter b_s is common to both components of the mixture model. To ensure that the membership parameter z_{is} is uniquely estimated for each component (to address the identifiability

Table 2. An illustration for displaying gain index.

	Gains			
Randomization	−Ve Gain	+Ve Gain	Proportion	Gain Index
Intervention	a	b	$p_2 = \frac{b}{(a+b)}$	$\eta = p_2 - p_1$
Comparison	c	d	$p_1 = \frac{d}{(c+d)}$	

problem), we constrained the components with $\mu_2 > \mu_1$. The average score for pupils with − ve gains is μ_1, and the average score pupils with positive gains is μ_2 and df is the degree of freedom. In addition, a common pooled variance σ^2 is assumed for both components in a multilevel model. The T-distribution is preferred to a normal distribution because of its thinner tail for smaller degrees of freedom (Kraemer & Paik, 1979; Lange et al., 1989). To reduce the proportion of overlapping data between the two components, we assumed two degrees of freedom as discussed by Evans and Erlandson (2004).

Given that we can now identify children that are likely to have a higher score post-intervention than pre-intervention (z_{is}) from a finite component mixture model, we calculated the percentages of children that are likely to make positive gain by cross-classifying the intervention group with the underlying latent construct for positive gain (z_{is}), as illustrated in Table 2. The Gain Index can, therefore, be presented in a 2 × 2 table in a similar way to the binomial effect size display (BESD). The Gain Index analytical framework is presented in Figure 3.

We implemented the model in R using R2jags package with vague priors. Although there are different options for vague priors, we have used common specifications implemented in WinBUGS and JAGS. The WinBUGS program used for the finite component mixture model is provided in the appendix. The priors for the average negative and positive gains were specified as $\mu_1 \sim N^{-}(0, 1000)$ and $\mu_2 \sim N^{+}(0, 1000)$. The priors for the inverse of variances are specified as $\frac{1}{\sigma^2} \sim \gamma(1000, 1000)$ and $\frac{1}{\sigma_b^2} \sim \gamma(1000, 1000)$. The membership parameters Z_{is} are sampled from a Bernoulli distribution $Z_{is} \sim Bern(p_1)$ where $p_1 \sim U(0, 1)$ and $p_2 = 1 - p_1$. The posterior estimates were generated using Gibb's sampling with 100000 iterations and burn-in of another 50000 iterations to train the Gibb's sampler. Figure A1 in the appendix provides the trace and density plots from MCMC for convergence check in each trial including Rhat. The Rhat estimate for each trial is 1, which suggests that the Bayesian model has converged.

Results

Our proposed first step in calculating the percentages of children that are more likely to have higher scores post-intervention than pre-intervention is to identify the underlying latent construct for positives and negative gains group using a finite component mixture model. One of the advantages of the mixture model is that the data determines the threshold between the positive and negative gain groups. Table 3 presents the component-specific parameters for the positive and the negative gain groups from a finite mixture model. The parameters p_2 and p_1 were directly estimated using the model specified in Equation (3), while GT and GC were derived from the same model as illustrated in Figure 3. In Trial 1, the average standardized gain score for the positive gain group was 0.63 (0.10, 1.27) and −1.00 (−1.65, −0.33) for the negative gain group. This means that children in the positive gain group were likely to make a progress of 0.63 standard deviations between the pre and post-intervention period, while those in the negative gain group on average declined by −1 standard deviation between the pre- and post-intervention period. In this trial, 54% of the children were likely to have higher scores post-intervention than pre-intervention whilst 46% of the children were likely to have a lower score post-intervention than pre-intervention. A similar pattern can be seen in Trial 2 where 39% of the children were likely to make positive gain with an average of 1.16 (0.57, 1.70) and 61% were likely to make a negative gain with an average of −0.61 (−1.15, −0.13) standard deviations. 37% of the children in Trial 3

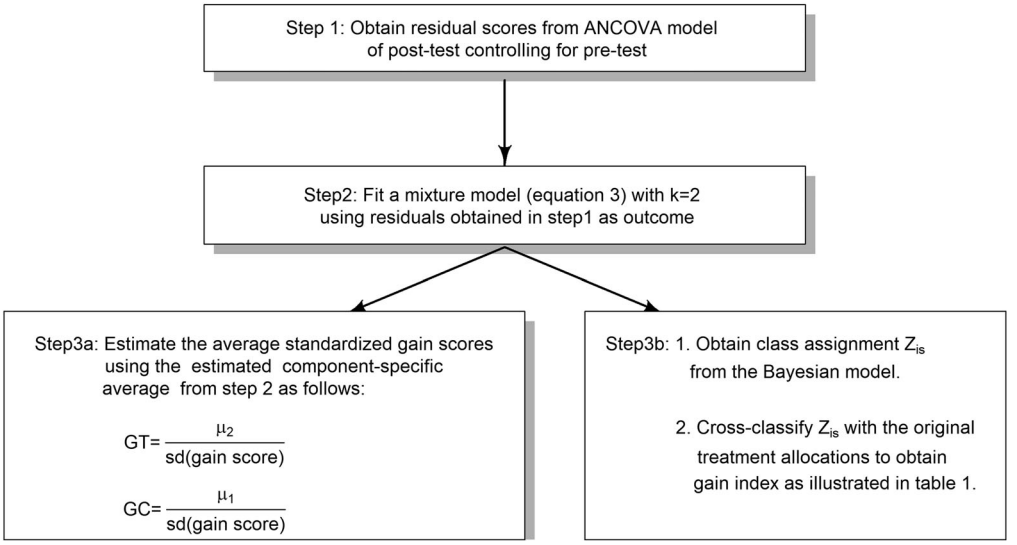

Figure 3. Flow diagram of finite mixture model analysis to obtain gain index.

Table 3. Component-specific parameter estimates from Bayesian shared parameter mixture model with 95% credible intervals in parentheses. GT is the standardized gain scores for the positive gain group, GC is the standardized gain scores for the negative gain group, p_2 is the proportion of children in the positive gain group and p_1 is the proportion of children in the negative gain group.

Parameters	Trial 1	Trial 2	Trial 3	Trial 4
GT	0.63	1.16	1.33	1.05
	(0.10, 1.27)	(0.57, 1.70)	(0.70, 1.87)	(1.05, 1.48)
GC	−1.00	−0.61	−0.91	−0.80
	(−1.65, −0.33)	(−1.15, −0.13)	(−1.33, −0.54)	(−1.16, −0.43)
p_2	0.54	0.39	0.37	0.42
	(0.34, 0.73)	(0.24, 0.56)	(0.22, 0.55)	(0.27, 0.56)
p_1	0.46	0.61	0.63	0.58
	(0.27, 0.66,)	(0.44, 0.76)	(0.45, 0.78)	(0.44, 0.73)

were likely to make positive gain with an average of 1.33 (0.70, 1.87) and 63% were likely to make negative gain with an average of −0.91 (−1.33, −0.54) standard deviations. Lastly, 42% of the children in Trial 4 were likely to make positive gains with an average of 1.05 (1.05, 1.48) and 58% of the children were likely to make negative gains with an average of −0.80 (−1.16, −0.43) standard deviations.

The results from the finite mixture model confirmed the expectation that some children will make progress between the pre- and post-intervention period whether they are in an intervention group or comparison group. However, one would expect the percentage of children in an intervention group that makes a positive gain to be higher than the percentage of children in the comparison group that were likely to make a positive gain, if the intervention is effective. The reverse pattern is expected for ineffective interventions or interventions with unintended consequences. Table 4 presents the percentages of children in the intervention and comparison groups that were likely to make positive or negative gains. The results presented in Table 4 were obtained as indicated in step 3b of Figure 3. In Trial 1, 60% of 149 children randomized to the intervention group were likely to make positive gain whilst 47% of 142 children in the comparison groups were likely to make positive gains. This means a difference of 13% with a 95% credible interval of 3% to 22% between the intervention and the comparison groups. In Trial 2, 45% of 159 children in the intervention group were likely to make positive gains compared to 33% of 143 children in the comparison group. This also shows that children in the intervention groups were more likely

Table 4. A 2 × 2 display of percentages of children in the intervention and comparison groups that were classified to have made negative or positive gains between pre and post-intervention periods.

Trials	Randomization	Gains		Proportion	Gain Index
		−Ve Gain	+Ve Gain		
Trial 1	Intervention	59	90	0.60	0.13 (0.03, 0.22)
	Comparison	75	67	0.47	
Trial 2	Intervention	88	71	0.45	0.12 (0.03, 0.21)
	Comparison	97	46	0.33	
Trial 3	Intervention	58	35	0.38	0.01 (−0.08, 0.12)
	Comparison	56	33	0.37	
Trial 4	Intervention	120	79	0.40	−0.04 (−0.12, 0.03)
	Comparison	106	86	0.44	

to make a positives gain than those in the comparison group with a percentage difference of 12% and 95% credible intervals of 3%–21%. In Trial 3, 38% of 93 children in the intervention group and 37% of 89 children in the comparison group were likely to make positive gains between the pre- and post-intervention period. This also means that children in the intervention group were more likely to make positive gains than those in the comparison group with a percentage difference of 1% and a credible interval of −8%–12%. However, in Trial 4, 40% of 199 children in the intervention group and 44% of 192 children in the comparison group were likely to make positive gains between the pre and post-intervention period. This is not an effective intervention as children in the comparison group were at least as likely as those in the intervention group to make positive gains. Figure 4 shows the posterior probability distribution of children in the positive gain group in each trial of the four representative trials. The posterior probabilities are plotted against residual of post-test scores obtained from the ANCOVA model. As expected, the higher the gain in test scores, the more likely it is for the pupils to make positive gain. Figure A2 in the appendix also shows how the distributions of children in the positive and negative gain groups are different.

We understand our proposed metric for assessing the impact of an intervention is not likely to be free from controversy because some researchers and statisticians consider reporting continuous scores as a binary outcome a bad practice, similar to the criticisms of BESD as it potentially loses data in representing the overall effects. However, there is also no benefit in reporting an arbitrary metric like an effect size, which is difficult to communicate to policymakers, parents, and schools, if it is not interpreted and understood correctly. It is interesting to note that our proposed method is strongly correlated with effect size (correlation = 0.92), although this would be expected given that the threshold between positive and negative gains are data dependent. Table A1 in the Appendix provides the estimated effect size and gain index, while Figure 5 shows the association between our proposed "gain index" and Hedges' effect size for eighteen trials, similarly funded by the EEF (Xiao et al., 2016).

Simulation study

A simulation study was conducted to investigate the performance of the finite component mixture model with varying number of pupils and schools. Assuming the average score for the − ve gain group is −1.89 with a residual variance of 35.05 and the variance of random effect is specified as 3.13, we simulated two components mixture model using Equation (3) as

$$f(y_{is}|b_s) = (1 - z_{is})\mathcal{T}(-1.89 + b_s, 35.05, 2) + z_{is}\mathcal{T}(\mu_2 + b_s, 35.05, 2). \quad (4)$$

where $z_{is} \sim Bern(0.5)$, $b_s \sim N(0, 3.13)$ and $\mu_2 = \emptyset * \sqrt{3.13 + 35.05}$. Note that $\emptyset = \{0.05, 1, 1.5, 2\}$ is the assumed standardized distance between the − ve gain and + ve gain group. We further assumed $N = \{10, 20\}$ to be the number of pupils per school and M = {10, 20, 30, 40, 50} to be the number of schools. For each combination of the parameters ($\emptyset$, n, m), 1000

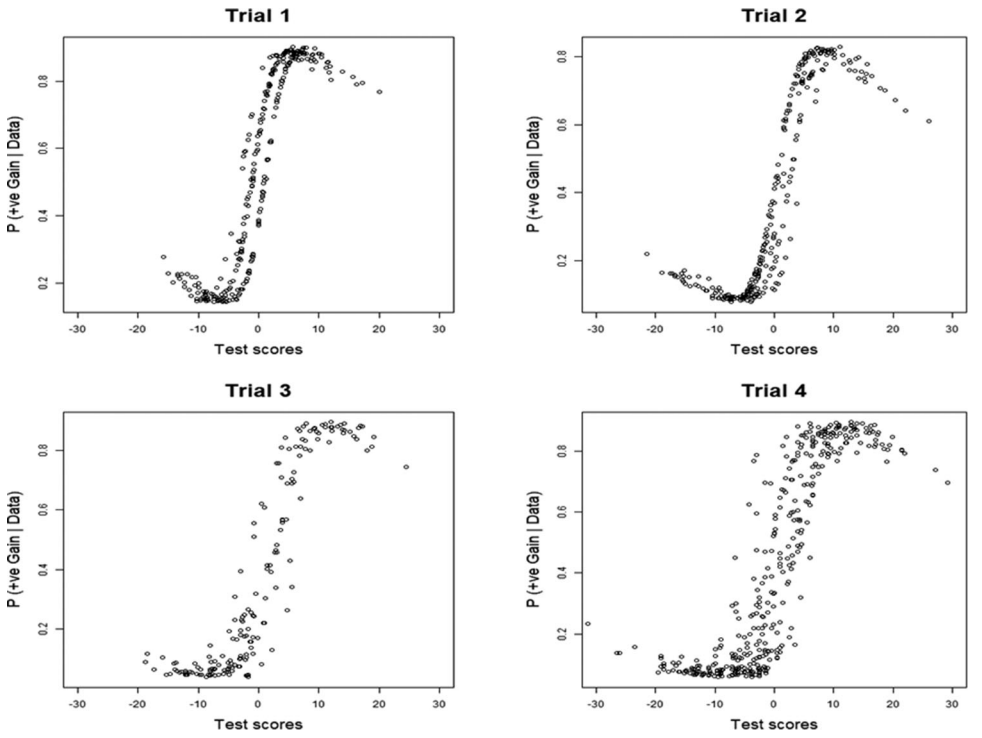

Figure 4. Distribution of posterior probability for positive gain group in each trial.

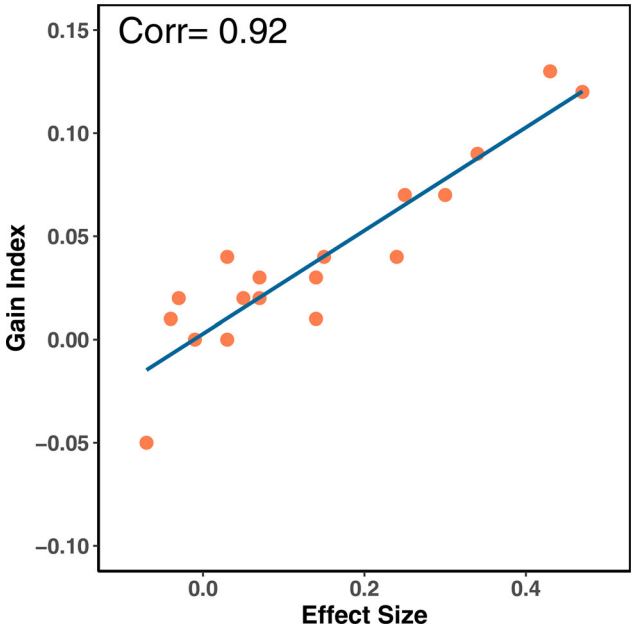

Figure 5. Strong correlation between the gain index and effect size commonly reported for educational interventions.

independent data were simulated from a t-distribution with two degrees of freedom. Note that the parameters for the simulation setting were informed by the data from Trial 1. The goal of the simulation study is to investigate the sensitivity and specificity of the proposed two components

Table 5. Investigation of classification accuracy for two-component mixture model using sensitivity and specificity.

		Standardized distance between − ve gain group and + gain group in standard deviation (SD)							
		0.05 SD		1 SD		1.5 SD		2 SD	
m	n	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
10	10	0.52 (0.42, 0.63)	0.52 (0.41, 0.63)	0.84 (0.69, 0.97)	0.85 (0.69, 0.97)	0.94 (0.84, 1.00)	0.94 (0.83, 1.00)	0.98 (0.92, 1.00)	0.98 (0.92, 1.00)
	20	0.52 (0.45, 0.60)	0.52 (0.45, 0.6)	0.85 (0.72, 0.95)	0.84 (0.72, 0.95)	0.94 (0.86, 0.99)	0.94 (0.86, 0.99)	0.98 (0.94, 1.00)	0.98 (0.94, 1.00)
20	10	0.52 (0.45, 0.60)	0.52 (0.45, 0.6)	0.84 (0.73, 0.95)	0.85 (0.73, 0.94)	0.94 (0.86, 0.99)	0.93 (0.86, 0.99)	0.98 (0.94, 1.00)	0.98 (0.94, 1.00)
	20	0.52 (0.47, 0.58)	0.52 (0.47, 0.58)	0.84 (0.76, 0.92)	0.84 (0.76, 0.92)	0.93 (0.89, 0.97)	0.93 (0.88, 0.97)	0.98 (0.95, 1.00)	0.98 (0.95, 0.99)
30	10	0.52 (0.47, 0.59)	0.52 (0.46, 0.59)	0.85 (0.73, 0.93)	0.84 (0.74, 0.93)	0.93 (0.88, 0.98)	0.93 (0.87, 0.98)	0.98 (0.95, 1.00)	0.98 (0.95, 1.00)
	20	0.52 (0.48, 0.56)	0.52 (0.48, 0.56)	0.84 (0.77, 0.91)	0.84 (0.76, 0.91)	0.93 (0.89, 0.97)	0.93 (0.9, 0.97)	0.98 (0.96, 0.99)	0.98 (0.96, 0.99)
40	10	0.52 (0.47, 0.57)	0.52 (0.47, 0.57)	0.84 (0.75, 0.93)	0.85 (0.75, 0.92)	0.94 (0.89, 0.97)	0.93 (0.89, 0.97)	0.98 (0.95, 0.99)	0.98 (0.95, 0.99)
	20	0.52 (0.48, 0.56)	0.52 (0.48, 0.56)	0.84 (0.78, 0.9)	0.84 (0.78, 0.9)	0.93 (0.9, 0.96)	0.93 (0.9, 0.96)	0.98 (0.96, 0.99)	0.98 (0.96, 0.99)
50	10	0.52 (0.47, 0.57)	0.52 (0.47, 0.57)	0.85 (0.76, 0.91)	0.84 (0.76, 0.91)	0.94 (0.89, 0.97)	0.93 (0.89, 0.97)	0.98 (0.95, 1.00)	0.98 (0.96, 1.00)
	20	0.52 (0.49, 0.55)	0.52 (0.49, 0.56)	0.84 (0.79, 0.89)	0.84 (0.79, 0.89)	0.93 (0.9, 0.96)	0.93 (0.9, 0.96)	0.98 (0.96, 0.99)	0.98 (0.96, 0.99)

mixture model. The sensitivity refers to "true positive" i.e. how well it classified children assign to the + ve gain group and specificity refers to "true negative" i.e. how well it classified children assigned to the − ve gain group.

Simulation results

The simulation results are presented in Table 5. As expected, the bigger the distance between the two groups the higher the specificity and the sensitivity of the mixture model. The sensitivity and specificity for a distance of one standard deviation (1 SD) is more than 80%. However, only about 50% of the children in the intervention and the comparison groups were correctly classified when the distance was 0.05 SD. This is expected since the distribution of the two components can be approximated due to the proximity of their individual distribution. Note that the distance between the two components in all the trials analyzed in this paper were bigger than 1 SD (See GT – GC in Table 3 for example).

Conclusion

The growing concern about null hypothesis significance testing as proof of effectiveness is only a part of the problem with quantifying evidence. A bigger issue is how evidence is communicated to educational stakeholders such as policymakers, parents and teachers. The current practice of reporting an effect size (standardized mean difference) and its associated confidence interval is a better approach for reporting findings from education trials than just a p-value. However, we know that an effect size is not easily understood by education stakeholders. This paper, therefore, proposed a pragmatic approach for analyzing education trials to improve communication and the interpretation of evidence for policymakers, parents, and schools. The gain index as proposed in this paper is not just a transformation of the effect size, but a more intuitive metric that relies on the latent patterns in the data. It relies on the expectation that an effective educational intervention will have at least two underlying distributions for the children with positive and negative gains. The proposed framework enables interpretation of the impact of an intervention in terms

of percentages of children that make progress between the pre and post-intervention period. It is an important metric that can help parents, teachers, and policymakers to understand that a positive finding from a randomized controlled education trial does not mean that every child benefits from the intervention. We have also shown that some children will make progress over time due to the normal activities in school or resulting from individual home support for the pupils. Some children may also benefit from an educational intervention beyond normal school progress. This work is therefore a positive step toward identifying more targeted or personalized support for struggling children, by identifying a group that appears to have progressed as a result of the intervention. It also has the potential to identify pupils that do not benefit from an intervention and facilitate empirical investigation of why an intervention may be effective for some pupils and not others.

The proposed gain index should also have implications for the economic analysis of educational interventions. The evaluation of the cost and benefit of an intervention based on the total number of pupils in the intervention group is likely to underestimate the true cost and benefit of the intervention. For an intervention that costs £1000 in a trial with 100 pupils in the intervention group, assuming zero cost for the comparison group, the crude cost per child is £10, but this is only the case if the intervention is equally beneficial for all the pupils in the intervention group. Suppose only 20 out of the 100 pupils in the intervention (20 percent) benefited from the intervention, then the actual cost per pupil who benefits is £50. Understanding that a positive finding from an intervention does not imply positive effects for all pupils is important for parents, teachers, and policymakers.

Lastly, the gain index is based on a Bayesian finite mixture model and its estimation process can be computationally intensive. However, the computational problem can be overcome by using suitable software such as R2jags and R2winbugs packages of R, Stata (Thompson et al., 2006) and SAS (Zhang et al., 2008) which can interact with Winbugs (Lunn et al., 2000). Winbugs is a statistical software package used for Bayesian analysis using Markov chain Monte Carlo methods. Furthermore, the proposed modeling approach can easily be implemented by adopting the freely available WinBUGS programme (MRC Biostatistics Unit, University of Cambridge, 2020). It can also be implemented using "proc mcmc" procedure in SAS statistical software. We have focused on the Bayesian finite mixture model with only two components because the primary interest of this study was in those pupils who made either positive or negative gain. However, it is possible to apply this method to more than two components. The direction set by our proposed method is mainly for understanding more clearly who might benefit from an intervention (and importantly who might not) which is important in terms of educational equity. We strongly believe that this paper is a positive addition to the existing literature on the estimation and communication of evidence in education.

Disclosure statement

None declared.

Funding

This research was funded by a grant to Durham University from the Education Endowment Foundation.

ORCID

Germaine Uwimpuhwe http://orcid.org/0000-0003-4122-7730
ZhiMin Xiao http://orcid.org/0000-0001-6464-2019

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology (London, England: 1953)*, *100*(Pt 3), 603–617. https://doi.org/10.1348/000712608X377117

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher*, *48*(4), 217–228. https://doi.org/10.3102/0013189X19848729

Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, *102*(1), 1–8. https://doi.org/10.1007/s10649-019-09908-4

Coe, R. (2002). *It's the effect size, stupid: What effect size is and why it is important.* http://www.leeds.ac.uk/educol/documents/00002182.html

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Revised edition. Academic press. https://doi.org/10.1086/ahr/88.3.760

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928. https://doi.org/10.1093/jpepsy/jsp004

Evans, R. B., & Erlandson, K. (2004). Robust Bayesian prediction of subject disease status and population prevalence using several similar diagnostic tests. *Statistics in Medicine*, *23*(14), 2227–2236. https://doi.org/10.1002/sim.1792

Ferguson, C. J. (2009). An effect size primer: a guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. https://doi.org/10.1037/a0015808

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Geoffrey, M., & Peel, D. (2000). *Finite mixture models*. Wiley.

Glass, G., Smith, M., & McGaw, B. (1981). *Meta-analysis in social science research*. Sage Publications.

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics & Gynaecology*, *125*(13), 1716–1716. https://doi.org/10.1111/1471-0528.15199

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Higgins, S., Katsipataki, M., Coleman, R., Henderson, P., Major, L., Coe, R., & Mason, D. (2015). *The Sutton Trust- Education endowment foundation teaching and learning toolkit*. Education Endowment Foundation.

Higgins, S. (2018). *Improving learning: Meta-analysis of intervention research in education*. Cambridge University Press.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hix-Small, H., Duncan, T. E., Duncan, S. C., & Okut, H. (2004). A multivariate associative finite growth mixture modeling approach examining adolescent alcohol and marijuana use. *Journal of Psychopathology and Behavioral Assessment*, *26*(4), 255–270. https://doi.org/10.1023/B:JOBA.0000045341.56296.fa

Hutchison, D., & Styles, B. (2010). *A guide to running randomised controlled trials for educational researchers*. NFER.

Izsák, A., & Jacobson, E. (2017). Preservice teachers' reasoning about relationships that are and are not proportional: A knowledge-in-pieces account. *Journal for Research in Mathematics Education*, *48*(3), 300–339.

Kraemer, H. C., & Paik, M. (1979). A central t approximation to the noncentral t distribution. *Technometrics*, *21*(3), 357–360. https://doi.org/10.2307/1267759

Kendall, J. (2003). Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal: EMJ*, *20*(2), 164–168. https://doi.org/10.1136/emj.20.2.164

Lange, K. L., Little, R. J., & Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, *84*(408), 881–896. https://doi.org/10.2307/2290063

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.

Lord, P., Bradshaw, S., Stevens, E., & Styles, B. (2015). *Perry beeches coaching programme: Evaluation report and executive summary*. Education Endowment Foundation.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337. https://doi.org/10.1023/A:1008929526011

Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., Clague, L., & Stiell, B. (2014). *Summer active reading programme: evaluation report and executive summary*. Education Endowment Foundation.
Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., & Clague, L. (2014). *Textnow transition programme*. Education Endowment Foundation.
McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, *71*(1), 173–180. https://doi.org/10.1111/1467-8624.00131
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361
Merrell, C., & Kasim, A. (2015). *Butterfly phonics: Evaluation report and executive summary*. Educational Endowment Foundation.
Miller, T. R., Hendrie, D., & Derzon, J. (2011). Exact method for computing absolute percent change in a dichotomous outcome from meta-analytic effect size: Improving impact and cost-outcome estimates. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *14*(1), 144–151. https://doi.org/10.1016/j.jval.2010.10.013
MRC Biostatistics Unit, University of Cambridge (2020). *Winbugs manual Volume 2*. Retrieved January 24, 2020, from https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/WinBUGS_Vol2.pdf.
Orylska, A., Hadwin, J., Kroemeke, A., & Sonuga-Barke, E. J. S. (2019). A growth mixture modelling study of learning trajectories in an extended computerised working memory training programme developed for young children diagnosed with attention-deficit/hyperactivity disorder. *Frontiers in Education*, *4*, 12. https://doi.org/10.3389/feduc.2019.00012
Randolph, J. J., & Edmondson, R. S. (2005). Using the binomial effect size display (BESD) to present the magnitude of effect sizes to the evaluation audience. *Practical Assessment, Research, and Evaluation*, *10*(1), 14.
Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, *45*(6), 775–777. https://doi.org/10.1037/0003-066X.45.6.775
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (Vol. 2). McGraw-Hill.
Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology*, *9*(5), 395–396. https://doi.org/10.1111/j.1559-1816.1979.tb02713.x
Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*(2), 166–169. https://doi.org/10.1037/0022-0663.74.2.166
Torgerson, C. J., & Torgerson, D. J. (2013). *Randomised trials in education: An introductory handbook*. Education Endowment Foundation.
Thompson, J., Palmer, T., & Moreno, S. (2006). Bayesian analysis in Stata with WinBUGS. *The Stata Journal: Promoting Communications on Statistics and Stata*, *6*(4), 530–549. https://doi.org/10.1177/1536867X0600600406
Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*(4), 473–481. https://doi.org/10.1037/0022-0167.51.4.473
Wang, M., & Bodner, T. E. (2007). Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, *10*(4), 635–656. https://doi.org/10.1177/1094428106289397
Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica*, *56*(3), 362–375. https://doi.org/10.1111/1467-9574.t01-1-00072
Xiao, Z., Kasim, A., & Higgins, S. (2016). Same difference? Understanding variation in the estimation of effect sizes from educational trials. *International Journal of Educational Research*, *77*, 1–14. https://doi.org/10.1016/j.ijer.2016.02.001
Zhang, Z., McArdle, J. J., Wang, L., & Hamagami, F. (2008). A SAS interface for Bayesian analysis with WinBUGS. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 705–728. https://doi.org/10.1080/10705510802339106